Migene: an Evidence-based Database of Genes and Phenotypes of Male Infertility

Taijiao Jiang (taijiao@ibms.pumc.edu.cn)
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China

Xiangzhe Guo
School of Basic Medical Sciences, Weifang Medical University, Weifang 261053, China

Fujun Peng
School of Basic Medical Sciences, Weifang Medical University, Weifang 261053, China

Tianrui Chen
Institute of Reproductive Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

Honglei Li
Suzhou Geneworks Biotechnology Co., Ltd., Suzhou 215123, China

Liting He
Institute of Reproductive Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

Lin Xiao
School of Basic Medical Sciences, Weifang Medical University, Weifang 261053, China

Xiuwei Ma
BaYi Children's Hospital, the Seventh Medical Center of PLA General Hospital, Beijing 100700, China

Songchen Dai
School of Basic Medical Sciences, Weifang Medical University, Weifang 261053, China

Shaoping Li
School of Basic Medical Sciences, Weifang Medical University, Weifang 261053, China

Shang Yuan
School of Basic Medical Sciences, Weifang Medical University, Weifang 261053, China

Tong Wang
School of Basic Medical Sciences, Weifang Medical University, Weifang 261053, China

Kai Zhao
Institute of Reproductive Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

Research Article

Keywords: male infertility, spermatogenic failure, mutation, gene-phenotypic association, enrichment analysis

Posted Date: January 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1216843/v1

Abstract

Male infertility (MI) is a disease with high heterogeneity. Its direct cause is spermatogenic failure, which lead to sperm abnormalities and quality deterioration in semen, and ultimately infertility, of which genetic factors account for about 30 percent. Although some knowledge bases had established the correlation between genetic factors and MI/spermatogenesis, they did not highlight the relationship between gene mutations and MI phenotypes, and even some had stopped updating. Meanwhile, a large mass of genotypes and phenotypes data were scattered and not effectively utilized with the wide application of high-throughput sequencing technology in MI research. To address these tough issues, a MIgene database (http://midb.geneworks.cn) was built through integrating existing data of gene-phenotypic MI from 989 literatures in PubMed and Medline database. A total of 22 information including 18 direct entries and 4 extended entries were extracted, filtered, and curated using bioinformatical and manual methods, resulting in 664 genes, 68 (standard) phenotypes (classified into 37 categories), 3606 mutants and 7985 studies in MIgene database. It was a web-accessible data repository and offered four search moles and six modules covering the genes, phenotypes, proteins, functions, homologous, cases, etc. Interestingly, many non-exonic variants could cause the MI, the same mutation could increase or decrease the MI in different phenotypes and races, the degree of gene-phenotypic association was presented by the enrichment analyses based on the principle of hypergeometric distribution. In general, MIgene not only had user-friendly interface for concise search, convenient browse, and customized downloads, but also provides early warning of disease risk and assist clinicians in timely diagnosis.

Introduction

Infertility is a global health problem among couples when they fail to conceive a child over one year of unprotected intercourse[1]. It occurs in approximately 15% of couples, half of which could be attributed to male infertility (MI) that makes a major threat to patients' family harmony, and mostly in developing countries[2,3]. The direct cause of male infertility is spermatogenesis failure, which lead to sperm abnormalities and quality deterioration in semen, and ultimately infertility, of which genetic factors account for about 30%[4–6].

Spermatogenesis was a sophisticated biological process responsible for development of spermatozoa from spermatogonia stem cells, and was elaborately regulated by multiple genes. Spermatogenesis failure usually presented the spermatogenic quantitative defects (azoospermia, oligozoospermia) and spermatogenic qualitative defects (globozoospermia, macrozoospermia)[5,7]. Currently, the known genetic abnormalities, including chromosome aberrations, Y chromosome microdeletion, epigenetics and post-transcriptional modification, sperm DNA damage, mitochondrial DNA (mtDNA) mutation, and genetic variants, had been identified the associations with idiopathic MI[8–11]. In this manuscript, it mainly underlined the relationship between variants of mtDNA and chromosome and MI.

Up to now, many efforts have been made by different researchers to build database or knowledgebase in various aspects of MI. GermOnline 4.0 was a cross-species database gateway focusing on high-throughput gene expression data related to germline development, the meiotic and mitosis cell cycle in normal or malignant cells[12]. SpermBase, a sperm RNA database, included the large and small sperm-borne RNA expression data for M.musculus, H.sapiens, etc[13]. GermlncRNA, a catalog of germ cell long non-coding RNAs, systematically annotated lncRNAs for each specialized germ cell stage using public annotations and Hybrid Transcriptome Assembly (HTA) approach[14]. SpermatogenesisOnline 1.0 used manual curation from 30, 233 articles published before 1 May 2012, which contained 1666 core genes and 762 extended genes that participated in spermatogenesis in 37 organisms[15]. ReproGenomics Viewer, a cross-species and cross-technology web-based resource of manually-curated sequencing datasets related to reproduction[16,17]. Besides, there are other databases for MI, such as GED[18], GermSAGE[19] and CFTR2[20]. However, these database or knowledgebases did not highlight the relationship between gene mutations and male infertility phenotypes and some have stopped updating. In recent years, with the wide application of high-throughput sequencing technology in MI research, more genes have been found and a large mass of data about clinical phenotypes also have been obtained. However, these data are relatively scattered and have no effective utilized. Therefore, it is important to integrate the existing data to construct a comprehensive, informative, and updatable database for genetic predispositions to MI, which could greatly facilitate the counseling, diagnosis, and therapy for MI.

In this study, MIgene, an evidence-based database of genes and phenotypes related to MI, was presented to fulfill the increasingly urgent need for data integration and resources. That 664 genes (515 genes from non-GWAS and 179 genes from GWAS), 3606 mutations containing SNPs (Single Nucleotide Polymorphisms), VNTRs (Variable Number of Tandem Repeats), and 68 phenotypes (37 categories) were collected from 989 articles. To the best of our knowledge, MIgene is the first genetic database for MI to conveniently browse, retrieve and download, which can facilitate to study the functions of MI for researchers and to provide the reference information for clinicians in prenatal diagnosis.

Methods

Literature search.

MIgene integrated genetic variants of MI from publications. Besides SNPs, other variants like VNTRs were also included. More than 200 combinations of different keywords were searched in the PubMed and Medline database (Supplementary Table S1), such as ‘aspermia AND mutation’, ‘spermatogenesis failure AND genomic alteration’, ‘severe oligozoospermia AND gene defects’, ‘spermatogenesis impairment AND various’, ‘oligozoospermia AND polymorphism’, ‘non-obstructive aspermia AND mutant’, ‘male sterile AND mutant’, ‘male infertility AND various’, ‘male infertility AND copy number’, ‘infertile men AND genetic alteration’, etc. These papers containing these keywords in the titles or abstracts were obtained and then were fetched through NCBI E-utilities API. By manual screening, the following papers such as reviews and research articles about pharmacology, sociology, electrophysiology, behavioral research, neurophysiology, chromosome aberrant, Y-chromosome microdeletions and cancer/tumor were excluded. In addition, the papers about non-human species and meta-analyses in this version were dropped off. Finally, the remaining literatures included in our data set of MIgene.

Data extraction, integration, and curation.

The full text of each eligible publication was downloaded and read carefully. The detailed information of each study was extracted manually by two or three researchers. The genetic, proteomic, clinical and demographic contents belonged to 18 direct entries were collected from papers and 4 extended entries (Supplementary Table S2), such as gene name, mutant, study type, clinical significance, phenotypes, and author comments, then, and were proofread, standardized, replenished, verified through bioinformatic and manual methods according to GeneBank[21], HGNC (Human Gene Nomenclature Committee)[22], dbSNP[23], GeneCards[24] and UniProt[25]. However, not all the mutations have their own rs_IDs, so variants without rs_IDs were encoded with Jrs000001, Jrs000002… (Supplementary Table S3). For phenotypes, many papers only contained the entry ‘spermatogenic failure’, but not ‘azoospermia’, ‘oligozoospermia’, or others. Therefore, MIgene only showed the raw data. These collected phenotypes were classified into 68 (standard) phenotypes and 37 categories by reference criteria established by the 5th WHO edition and different forms of writing or synonyms (Table 1 and Supplementary Table S4-S6)[26]. The concepts of phenotypes were given by HPO[27], OMIM[28], Wikipedia (https://en.wikipedia.org) and the 5th edition of the WHO Laboratory Manual[26].

Table 1
Categories of main phenotypes related to male infertility.

Phenotypes categories	Phenotypes (standardization)	No. of studies
azoospermia	azoospermia, non-obstructive azoospermia	46.47% (3711/7985)
spermatogenic failure	spermatogenic failure, hypospermatogenesis, sertoli cell-only syndrome, meiotic arrest	45.37% (3623/7985)
oligozoospermia	oligozoospermia, extreme oligozoospermia, severe oligozoospermia, moderate or mild oligozoospermia	34.64% (2766/7985)
obstructive azoospermia	obstructive azoospermia	12.57% (1004/7985)
CAVD	CAVD, CBAVD, CUAVD	12.44% (993/7985)
asthenospermia	asthenospermia, severe asthenospermia	11.53% (921/7985)
teratospermia	teratospermia, macrozoospermia, globozoospermia, acephalic spermatozoa syndrome, MMAF	7.64% (610/7985)
oligo-astheno-teratozoospermia	oligo-astheno-teratozoospermia	5.96% (476/7985)
oligoasthenozoospermia	oligoasthenozoospermia, severe oligoasthenozoospermia, moderate oligoasthenozoospermia	5.28% (422/7985)
normozoospermia	normozoospermia	4.46% (356/7985)
disorders of sex development	male ambiguous genitalia, anorchia, cryptorchidism, disorders of sex development, virilization, male pseudohermaphroditism, testicular dysgenesis syndrome, delayed puberty, Leydig cell hyperplasia, gonadal dysgenesis, gynaecomastia, aplasia of the epididymis, hypospadias	3.18% (254/7985)
non-normozoospermia	non-normozoospermia	2.81% (224/7985)
male infertility (only one phenotype)	male infertility (only one phenotype)	2.48% (198/7985)
spermatogenesis maturation arrest	spermatogenesis maturation arrest	1.94% (155/7985)
gonadotropin deficiency	hypergonadotropic hypogonadism, male hypogonadotropic hypogonadism	1.39% (111/7985)
androgen insensitivity syndrome	androgen insensitivity syndrome, complete androgen insensitivity syndrome, mild androgen insensitivity syndrome, partial androgen insensitivity syndrome	1.20% (96/7985)
polycystic kidney disease	polycystic kidney disease	1.19% (95/7985)
asthenoteratozoospermia	asthenoteratozoospermia	1.09% (87/7985)
oligoteratozoospermia	oligoteratozoospermia	0.45% (36/7985)
varicocele	varicocele	0.41% (33/7985)

To illustrate the relationship between candidate genes and MI, statistical results were classified into 'related-damage', 'related-protection', 'unrelated' and 'unknown' according to their statistical evidence in the original publications. The results with $p < 0.05$ for non-GWAS or $p < 1x\ 10^{-8}$ for GWAS usually were defined as 'related' unless the authors suggested some other values or the associated studies according to the references[29]. However, many mutants produced the opposite consequence. For instance, the catalase C262T polymorphism indicated that CAT-262T/T genotype conferred less susceptibility to MI[30]. Hence, the 'related' was divided into two classes: 'related-damage' represented the results related to the increase of the risk of MI and 'related-protection' represented the results related to the decrease of the risk of MI. The 'unrelated' represented the results with $p > 0.05$ for non-GWAS or $p > 1x10^{-5}$ for GWAS. For 'unknown' results, the values were below these thresholds of GWAS or the original papers did not provide the clinical significance. If other statistical values were used, the criteria would be referred to as the statistical method in original papers. All the clinical results were checked by more than two researchers, the opposite consequences were verified after discussion.

Function annotations.

To further understand the function of all the genes associated with MI, extensive functional knowledge and data from the online database were gathered. Protein expression levels of subcellular-location were retrieved from Compartment. The position and function of the peptides and proteins were annotated using UniProt[25]. In addition, other information was provided, such as co-expression protein, protein-protein interactions, and enriched functional pathways from GeneCards[24], String[31], GO (Gene Ontology)[32] and KEGG[33], respectively.

Enrichment analysis of genes and phenotypes.

For the enrichment analysis of the association between gene and phenotype, algorithm of hypergeometric distribution[34,35] was applied. The results were named the enrichment scores of gene or phenotype enrichment analysis (Supplementary Table S7-S8). The -lg (p-value) was calculated with the enrichment score plus 0.0001 (this value could be random given), then logarithms and minus sign. After searching for a gene, the phenotype enrichment results can be acquired (Supplementary Table S7). By using the same strategy, after searching for a phenotype, the results of gene enrichment could be obtained (Supplementary Table S8).

Web interface configuration.

MIgene was established as an integrated information resource, in which the whole data were stored and managed in a MySQL relational database and implemented using node.js, JavaScript, vue and egg.js. They are platform independent, open and free source software and support multi-user to browse the web. The web interface is available online at http://midb.geneworks.cn/introduce.

Results

Data collection and curation.

Refer to the mentioned workflow of "Materials and methods" (Figure 1), we screened and selected 989 literatures related to MI from 25,312 papers, then gave 22 different entries to describe the patient. After processing these data, that 664 genes (515 genes from non-GWAS and 179 genes from GWAS), 68 phenotypes, 3606 mutants and 7985 studies were contained in MIgene (Figure 1 and Supplementary Table S9-S10).

Spermatogenesis is a highly organized process of cell proliferation in seminiferous tubules and terminal differentiation for the development of mature spermatozoa. If spermatogenesis is disturbed, it will cause azoospermia, oligozoospermia and other

defects of sperm count, motility, and morphology[10,36]. Among the whole phenotypes, the azoospermia, spermatogenic failure, and oligozoospermia studies account for 46.47%, 45.37%, 34.64%, respectively.

Analysis of genes, molecular consequence, variant types for clinical significance.

From 989 papers about MI, a total of 664 genes were obtained and classified into four clinical significance and two study types (Figure 2A, Table 2 and Supplementary Table S11-S12). Among these genes, there were 515 genes from non-GWAS, 179 genes from GWAS and 30 genes coexisting in them. Besides, there were 280 genes associated with more than two types of clinical significance. For example, the c.2039A>G mutant of FSHR gene showed four types of clinical significance under different conditions including phenotypes, zygosity and ethnicity, etc (Table 3)[37,38], which suggested the same variants of one gene performed different clinically statistics results. Distribution of clinical significance of all the genes and variants were summarized (Supplementary Table S13), which suggested that MI is a multifactorial disease.

Table 2
Distribution of genes among clinical significance for study types.

Study type Clinical significance	Sum	non-GWAS	GWAS	Both studies
related-damage	363	334	46	17
related-protection	47	46	1	0
unrelated	534	410	141	17
unknown	76	70	6	0
Total	664	515	179	30

Table 3
The partial results of rs6166 c.2039A>G (p. Asn680Ser) of FSHR gene.

Studies ID	PMID	Position	Zygosity	Clinical significance	Design type	Study type	Population origin	Phenotypes
S000658	20454649	48962782	heterozygous	related-damage	case-control	non-GWAS	Turkish	spermatogenesis impairment, male infertility: non-obstructive azoospermia
S000659	20454649	48962782	heterozygous	related-protection	case-control	non-GWAS	Turkish	spermatogenesis impairment, male infertility: severe oligozoospermia
S000660	20454649	48962782	homozygous	related-protection	case-control	non-GWAS	Turkish	spermatogenesis impairment, male infertility: non-obstructive azoospermia
S000661	20454649	48962782	homozygous	unrelated	case-control	non-GWAS	Turkish	spermatogenesis impairment, male infertility: severe oligozoospermia
S000665	10022448	48962782	heterozygous	unrelated	case-control	non-GWAS	German	male infertility: non-obstructive azoospermia, severe oligozoospermia
S000666	10022448	48962782	homozygous	unrelated	case-control	non-GWAS	German	male infertility: non-obstructive azoospermia, severe oligozoospermia
S002696	17169197	48962782	heterozygous	unrelated	case-control	non-GWAS	Italian	spermatogenesis impairment, male infertility: non-obstructive azoospermia
S002702	17169197	48962782	heterozygous	unrelated	case-control	non-GWAS	Italian	spermatogenesis impairment, male infertility: severe oligozoospermia

Fortunately, there were 103 genes (non-GWAS: 85 genes, GWAS: 19 genes) exclusively in "related-damage" patients and corresponded to 38 phenotypes, the genes' number for which phenotypes was counted and found that the top three phenotypes were spermatogenic failure (59 genes), azoospermia (47 genes), asthenospermia (20 genes) (Figure 2B and Supplementary Table S14).

Further, that 37.5% missense, 19.1% intron and 10.2% synonymous variants were the top three molecular consequence (Figure 2C) in MIgene. In the related-damage group, the top three results were 44.3% missense, 10.1% intron and 9.8% splice site (Figure 2D). Notablely, the intron mutations could affect MI in accordance with intron retention has the extent and functional significance[39].

The comprehensive collection of MIgene database allowed us to have an overview of related-damage genes among different chromosome. The gene ontology analysis revealed that every chromosome had a certain number of genes except chromosome 21 (Figure 2E and Supplementary Table S15). Importantly, a lot of mtDNA genes participated in MI. For example, the mtDNA 4977 deletion was found to be related to MI[40,41].

Enrichment analysis of genes and phenotypes.

To find further evidence for the association between genes and phenotypes, an enrichment analysis was performed on the basis of the principle of the hypergeometric distribution. The enrichment results interpreted that the larger the number of samples was for the enriched item in the database, the more stable the results of enrichment were (Figure 3). To take oligoteratozoospermia as an example, in related-damage group, the prioritization for MI candidate genes is presented in Figure 3A. We obtained the most relevant gene PLOG with oligoteratozoospermia.

It is well known that one gene could generate the different phenotypes, thus the phenotype enrichment rank for the gene was further explored. By using PLOG as a training gene, the phenotype enrichment analysis was ranked in graphics (Figure 3B). The top of these consequences was oligoteratozoospermia phenotypes in accord with Figure 3A.

Data search and navigation.

MIgene provides users a powerful and multi-faceted search engine and a user-friendly interface to access, browse and retrieve different data types and analysis results. The website interface comprises seven sections including “Home”, “Browser”, “Submit case”, “Download”, “Tutorial”, “Contact” and “Analysis” (Figure 4). On the “Home” page, a brief introduction of MI, information accessible in the database and gene or phenotype search are provided. There are four search modules, ‘Gene Symbol’, ‘Phenotype’, ‘rs_ID’ and ‘Mutant’. Furthermore, these symbols are not only auto-completed after typing some letters in their corresponding search box, but also cross-accessed using inter-linkages. After selecting “Browser” in the navigation bar, the complete list of MI including genes, related phenotypes, clinical significance, and supporting evidence, could be randomly browsed. On the “Submit case” page, the users could submit new genes, mutants, and phenotypes to our database. These data will be stored, curated and then entered the database. At the same time, this MIgene database will be updated periodically according to the latest publications. The ‘Tutorial’ page presents the database’s guidelines.

MIgene provides a detailed report for each gene. Firstly, to take the gene FSHR as an example, MIgene showed basic information of gene and protein including protein sequence annotations, function analysis and related external databases such as OMIM[28], InterPro[42], KEGG[33], GO[32], String[31], Compartment, etc. In addition, homology, enrichment phenotypes, and co-expression proteins were also obtained. For variant information of FSHR, the users of MIgene could not only get the variant types and its statistical results but also download the filtered contents at any moment. After the “view” button was clicked, the whole detailed information would be displayed for this genomic mutation. Also, the number of phenotypes and clinical significance associated with the gene was counted respectively. For example, the oligozoospermia was one of the phenotypes related to FSHR. There were 109 studies about it, which were divided into four groups: 5 of related-damage, 5 of related-protection, 97 of unrelated and 2 of unknown. Secondly, for phenotype, MIgene defined the phenotype, the number of studies, the information of enrichment genes and other contents including SNPs, indel, deletion, duplication, insertion, and related clinical significance. Thirdly, in rs_ID modules, the basic information of rs_ID, the number of studies and statistical clinical significance were exhibited by MIgene. Finally, this database provided a powerful and convenient way to search for the mutants of genes and phenotypes for MI.

Discussion

As a comprehensive and first genetic database of MI, MIgene included many genes, clinical phenotypes, and basic information of the patients. Authors have tried the best to extract the information through in-depth reading manually. At last, 664 genes (non-GWAS: 515 genes, GWAS: 179 genes), 68 phenotypes (37 categories), 3606 mutants and 7985 studies were obtained from 989 published articles. By integrating these data, MIgene was developed to provide a panoramic view of the current genetic research of MI and the association between genotype and phenotypes. MIgene aimed to act as not only an integrated genetic resource for MI but also as a flexible application platform for prenatal diagnosis' results for MI.

Several recent studies have demonstrated intron retention is a central component regulating gene expression during normal development as well as stress response and disease[39,43]. About 20% of non-exonic variants mainly included intron, splice site participated in the pathogenesis of MI. For example, the ACE I/D mutant in intron 16 is associated with abnormal seminal variables causing MI[44]. The variant DNAH1 c.8626-1G>A in intron 54 causes a frameshift mutation in the new transcript to induce a premature stop codon, which results in multiple morphological abnormalities of the sperm flagella (MMAF) in semen and then leads to MI [45]. In the non-exonic variants of related-damage subsets, 288 mutations in 141 genes causing 41 phenotypes (24 categories) were obtained from 245 published papers, some of which need to be focused on in the further studies.

In addition, several issues should be considered for MIgene. First, there is currently no universal, rapid and efficient system to screen the full text instead of manual screening[46,47]. It seriously slows down the update, especially when massive data of genes and mutants are being produced along with the development of new technologies. Therefore, the automatic mining methods should be exploited and used for updating male infertility-related data in the future. Second, the curation of clinical significance only confirmed according to the publications, but the supporting data in the publications are not validated by ourselves, which may lead to partial false positive results. In the next few years, the issues mentioned above are expected to be concerned and solved in future versions.

In summary, MI is a complicated disease that can be influenced by multi-factors including genes, phenotypes, mutation types, genetic background, ethnicity, environment and even zygosity. However, the MIgene database were established, which are publicly available for researchers and clinicians. Furthermore, it elucidates the association between genetypes and phenotypes (especially spermatogenic failure). This study could help users understand the complex biological process and mechanisms of MI, and provides references to the prenatal diagnosis' results.

Declarations

Acknowledgements

This work was supported by Central Public-Interest Scientific Institution Basal Research Fund (2017PT31026, 2018PT31016), National Basic Research Program of China (82103571), Natural Science Foundation of Shandong Province (ZR202103040133, ZR2021QH295), Weifang Medical College Doctoral Start Fund (2020BSQD40), Innovation Training Program for College Students in Shandong Province (X2021003X2021198).

Author Contributions

X.G., T.C., L.H., S.D., S.L. and S.Y. conducted the papers collection and filtration. X.G., F.P., T.W., T.C., and L.X. abstracted the papers contents. F.P. and K.Z. designed the MIgene website. H.L. developed the MIgene website and collected partial web data. X.G., F.P., and T.W. obtained and analyzed the MIgene dataset and wrote original manuscript. X.M., K.Z. and T.J. provided the advices and guidance for website design and paper writing. T.W., K.Z. and T.J. were responsible for the manuscript revision.

Conflict of interest

The authors have declared that no conflict of interests.

Data availability

All data generated or analysed during this study are included in this article, its Supplementary Tables and its MIgene database (http://midb.geneworks.cn/download). And, all experimental protocols were performed in accordance with relevant guidelines.

References

1. Moraes, A. P. *et al.* Incidence and main causes of severe maternal morbidity in Sao Luis, Maranhao, Brazil: a longitudinal study. *Sao Paulo Med J* **129**, 146–152 (2011).
2. Sharlip, I. D. *et al.* Best practice policies for male infertility. *Fertil Steril* **77**, 873–882 (2002).
3. Agarwal, A., Mulgund, A., Hamada, A. & Chyatte, M. R. A unique view on male infertility around the globe. *Reprod Biol Endocrinol* **13**, 37, doi:10.1186/s12958-015-0032-1 (2015).
4. Asero, P. *et al.* Relevance of genetic investigation in male infertility. *J Endocrinol Invest* **37**, 415–427, doi:10.1007/s40618-014-0053-1 (2014).
5. Krausz, C. & Riera-Escamilla, A. Genetics of male infertility. *Nat Rev Urol* **15**, 369–384, doi:10.1038/s41585-018-0003-3 (2018).
6. Jafari, H., Mirzaiinajmabadi, K., Roudsari, R. L. & Rakhshkhorshid, M. The factors affecting male infertility: A systematic review. *Int J Reprod Biomed* **19**, 681–688, doi:10.18502/ijrm.v19i8.9615 (2021).
7. Cannarella, R., Condorelli, R. A., Mongioi, L. M., La Vignera, S. & Calogero, A. E. Molecular Biology of Spermatogenesis: Novel Targets of Apparently Idiopathic Male Infertility. *Int J Mol Sci* **21**, doi:10.3390/ijms21051728 (2020).
8. O'Flynn O'Brien, K. L., Varghese, A. C. & Agarwal, A. The genetic causes of male factor infertility: a review. *Fertil Steril* **93**, 1–12, doi:10.1016/j.fertnstert.2009.10.045 (2010).
9. Bracke, A., Peeters, K., Punjabi, U., Hoogewijs, D. & Dewilde, S. A search for molecular mechanisms underlying male idiopathic infertility. *Reprod Biomed Online* **36**, 327–339, doi:10.1016/j.rbmo.2017.12.005 (2018).
10. Cannarella, R., Condorelli, R. A., Duca, Y., La Vignera, S. & Calogero, A. E. New insights into the genetics of spermatogenic failure: a review of the literature. *Hum Genet* **138**, 125–140, doi:10.1007/s00439-019-01974-1 (2019).
11. Gunes, S. & Esteves, S. C. Role of genetics and epigenetics in male infertility. *Andrologia* **53**, e13586, doi:10.1111/and.13586 (2021).
12. Lardenois, A., Gattiker, A., Collin, O., Chalmel, F. & Primig, M. GermOnline 4.0 is a genomics gateway for germline development, meiosis and the mitotic cell cycle. *Database (Oxford)* 2010, baq030, doi:10.1093/database/baq030 (2010).
13. Schuster, A. *et al.* SpermBase: A Database for Sperm-Borne RNA Contents. *Biol Reprod* **95**, 99, doi:10.1095/biolreprod.116.142190 (2016).
14. Luk, A. C. *et al.* GermlncRNA: a unique catalogue of long non-coding RNAs and associated regulations in male germ cell development. *Database (Oxford)* 2015, bav044, doi:10.1093/database/bav044 (2015).
15. Zhang, Y. *et al.* SpermatogenesisOnline 1.0: a resource for spermatogenesis based on manual literature curation and genome-wide data mining. *Nucleic Acids Res* **41**, D1055-1062, doi:10.1093/nar/gks1186 (2013).
16. Darde, T. A. *et al.* The ReproGenomics Viewer: an integrative cross-species toolbox for the reproductive science community. *Nucleic Acids Res* **43**, W109-116, doi:10.1093/nar/gkv345 (2015).
17. Darde, T. A. *et al.* The ReproGenomics Viewer: a multi-omics and cross-species resource compatible with single-cell studies for the reproductive science community. *Bioinformatics* **35**, 3133–3139, doi:10.1093/bioinformatics/btz047 (2019).
18. Bai, W. *et al.* GED: a manually curated comprehensive resource for epigenetic modification of gametogenesis. *Brief Bioinform* **18**, 98–104, doi:10.1093/bib/bbw007 (2017).

19. Lee, T. L. *et al.* GermSAGE: a comprehensive SAGE database for transcript discovery on male germ cell development. *Nucleic Acids Res* **37**, D891-897, doi:10.1093/nar/gkn644 (2009).
20. Castellani, C. & team, C. CFTR2: How will it help care? *Paediatr Respir Rev* **14 Suppl 1**, 2–5, doi:10.1016/j.prrv.2013.01.006 (2013).
21. Brown, G. R. *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res* **43**, D36-42, doi:10.1093/nar/gku1055 (2015).
22. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res* **43**, D1079-1085, doi:10.1093/nar/gku1071 (2015).
23. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–311, doi:10.1093/nar/29.1.308 (2001).
24. Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database (Oxford)* 2010, baq020, doi:10.1093/database/baq020 (2010).
25. UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212, doi:10.1093/nar/gku989 (2015).
26. Ford, W. C. Comments on the release of the 5th edition of the WHO Laboratory Manual for the Examination and Processing of Human Semen. *Asian J Androl* **12**, 59–63, doi:10.1038/aja.2008.57 (2010).
27. Kohler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res* **47**, D1018-D1027, doi:10.1093/nar/gky1105 (2019).
28. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**, D789-798, doi:10.1093/nar/gku1205 (2015).
29. Zhang, L. *et al.* ADHDgene: a genetic database for attention deficit hyperactivity disorder. *Nucleic Acids Res* **40**, D1003-1009, doi:10.1093/nar/gkr992 (2012).
30. Sabouhi, S., Salehi, Z., Bahadori, M. H. & Mahdavi, M. Human catalase gene polymorphism (CAT C-262T) and risk of male infertility. *Andrologia* **47**, 97–101, doi:10.1111/and.12228 (2015).
31. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**, D561-568, doi:10.1093/nar/gkq973 (2011).
32. Gene Ontology, C. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43**, D1049-1056, doi:10.1093/nar/gku1179 (2015).
33. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109-114, doi:10.1093/nar/gkr988 (2012).
34. Kotaru, A. R., Shameer, K., Sundaramurthy, P. & Joshi, R. C. An improved hypergeometric probability method for identification of functionally linked proteins using phylogenetic profiles. *Bioinformation* **9**, 368–374, doi:10.6026/97320630009368 (2013).
35. Zhang, Y. *et al.* Integrated analysis of mutation data from various sources identifies key genes and signaling pathways in hepatocellular carcinoma. *PLoS One* **9**, e100854, doi:10.1371/journal.pone.0100854 (2014).
36. Neto, F. T., Bach, P. V., Najari, B. B., Li, P. S. & Goldstein, M. Genetics of Male Infertility. *Curr Urol Rep* **17**, 70, doi:10.1007/s11934-016-0627-x (2016).
37. Balkan, M. *et al.* FSHR single nucleotide polymorphism frequencies in proven fathers and infertile men in Southeast Turkey. *J Biomed Biotechnol* 2010, 640318, doi:10.1155/2010/640318 (2010).
38. Collodel, G. *et al.* Alterations of the FSH and LH receptor genes and evaluation of sperm ultrastructure in men with idiopathic hypergonadotropic hypogonadism. *J Assist Reprod Genet* **30**, 1101–1108, doi:10.1007/s10815-013-0055-5 (2013).
39. Jacob, A. G. & Smith, C. W. J. Intron retention as a component of regulated gene expression programs. *Hum Genet* **136**, 1043–1057, doi:10.1007/s00439-017-1791-x (2017).

40. Kao, S., Chao, H. T. & Wei, Y. H. Mitochondrial deoxyribonucleic acid 4977-bp deletion is associated with diminished fertility and motility of human sperm. *Biol Reprod* **52**, 729–736, doi:10.1095/biolreprod52.4.729 (1995).
41. Dhillon, V. S., Shahid, M. & Husain, S. A. Associations of MTHFR DNMT3b 4977 bp deletion in mtDNA and GSTM1 deletion, and aberrant CpG island hypermethylation of GSTM1 in non-obstructive infertility in Indian men. *Mol Hum Reprod* **13**, 213–222, doi:10.1093/molehr/gal118 (2007).
42. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* **43**, D213-221, doi:10.1093/nar/gku1243 (2015).
43. Wong, J. J., Au, A. Y., Ritchie, W. & Rasko, J. E. Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. *Bioessays* **38**, 41–49, doi:10.1002/bies.201500117 (2016).
44. Zalata, A. A., Morsy, H. K., Badawy Ael, N., Elhanbly, S. & Mostafa, T. ACE gene insertion/deletion polymorphism seminal associations in infertile men. *J Urol* **187**, 1776–1780, doi:10.1016/j.juro.2011.12.076 (2012).
45. Amiri-Yekta, A. *et al.* Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new DNAH1 mutations. *Hum Reprod* **31**, 2872–2880, doi:10.1093/humrep/dew262 (2016).
46. Ongenaert, M. *et al.* PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res* **36**, D842-846, doi:10.1093/nar/gkm788 (2008).
47. Van Noorden, R. Trouble at the text mine. *Nature* **483**, 134–135, doi:10.1038/483134a (2012).

Figures

Figure 1. Process of data collection and curation in MIgene.

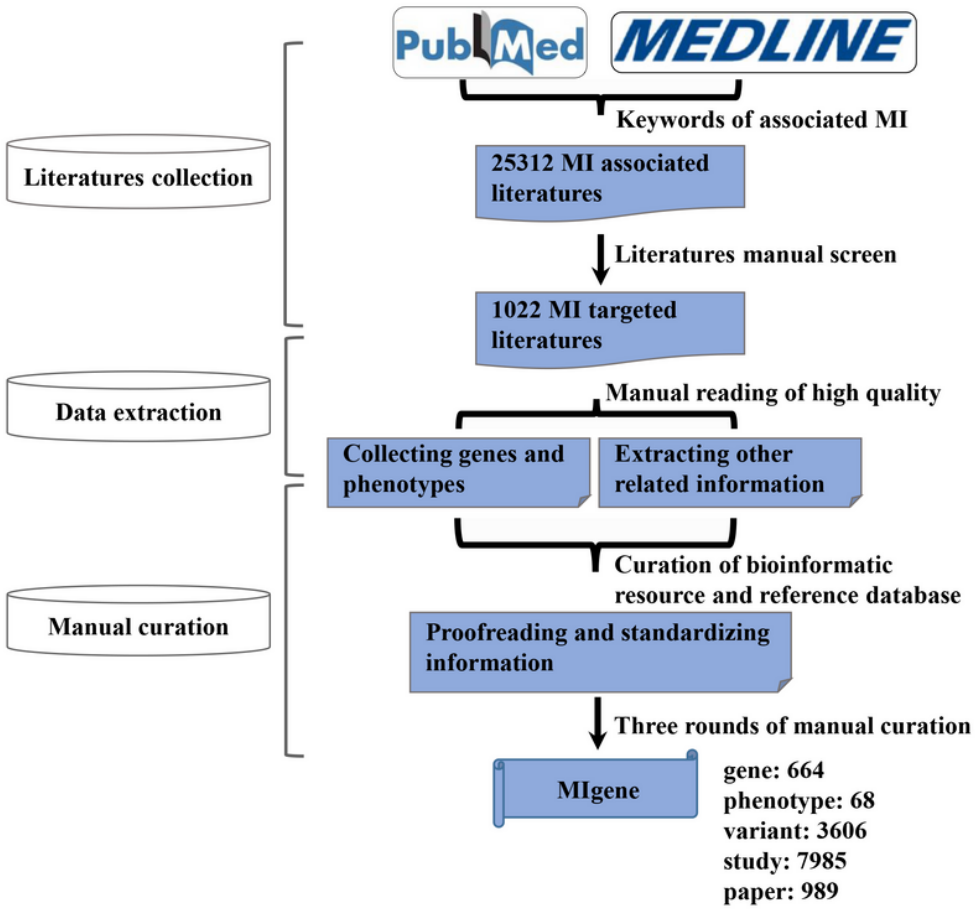

Figure 1

Process of data collection and curation in MIgene. The workflow of this project is divided into three sections, literature collection, data extraction and manual curation. MI: male infertility.

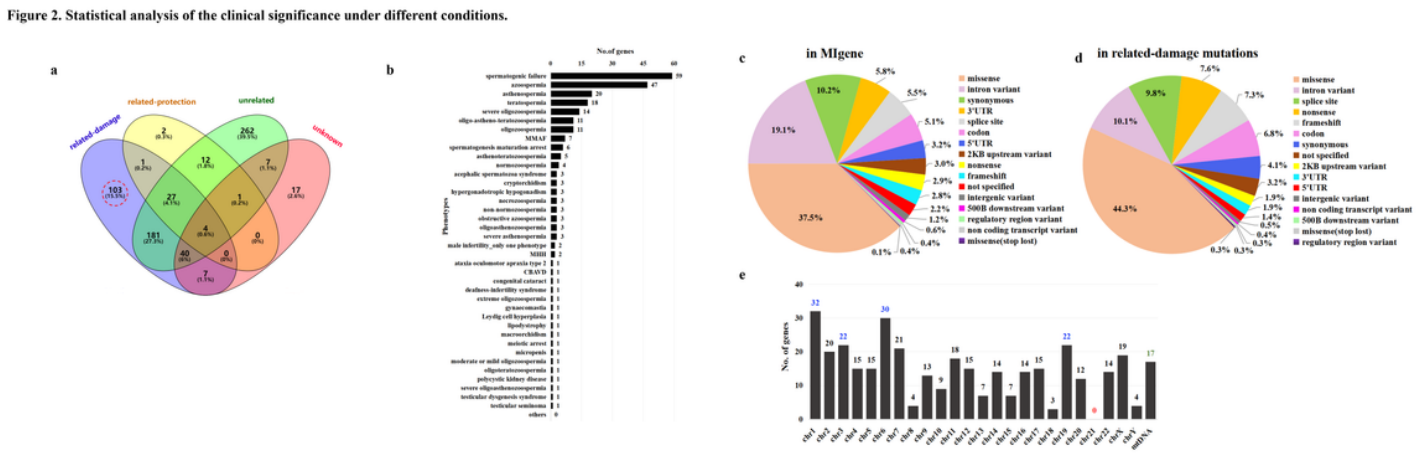

Figure 2

Statistical analysis of the clinical significance under different conditions. (a) Comparison of genes among clinical significance. (**b)** Statistics of genes corresponding different phenotypes in only 103 related-damage genes. (**c)** Statistics of molecular consequence for all the mutants in MIgene. The top three of molecular consequence are missense, intron and synonymous variants. (**d)** Distribution of molecular consequence for the related-damage events in MIgene. The missense, intron variants and splice site are the top three. (**e)** Analysis of related-damage genes for different chromosome. The related-damage genes mainly locate in chr1, chr6 and chr3 and chr19 besides mtDNA. Interestingly, there is no related-damage gene in chr21.

MMAF: multiple morphological abnormalities of the sperm flagella, MHH: male hypogonadotropic hypogonadism, CAVD: congenital absence of the vas deferens, CBAVD: congenital bilateral absence of the vas deferens.

Figure 3. Enrichment analysis of association between genes and phenotypes.

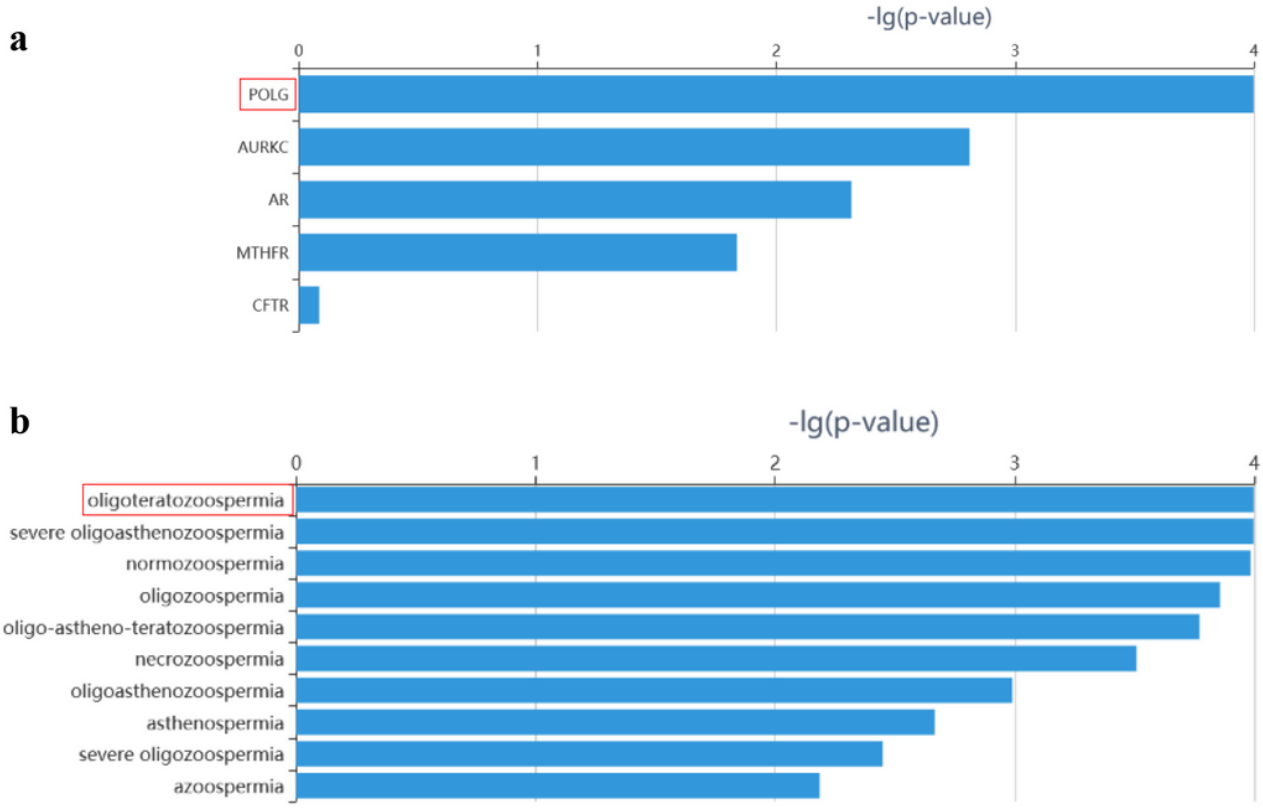

Figure 3

Enrichment analysis of association between genes and phenotypes. (a) the related-damage genes by taking oligoteratozoospermia as example. (b) The prioritization of phenotypes for related-damage by taking FOLG gene for example. The -lg (p-value) equals to -lg (enrichment score + 0.0001). The whole data sets of -lg (p-value) are displayed in Supplementary Table S7-S8.

Figure 4. The functional modules and screenshot of MIgene database.

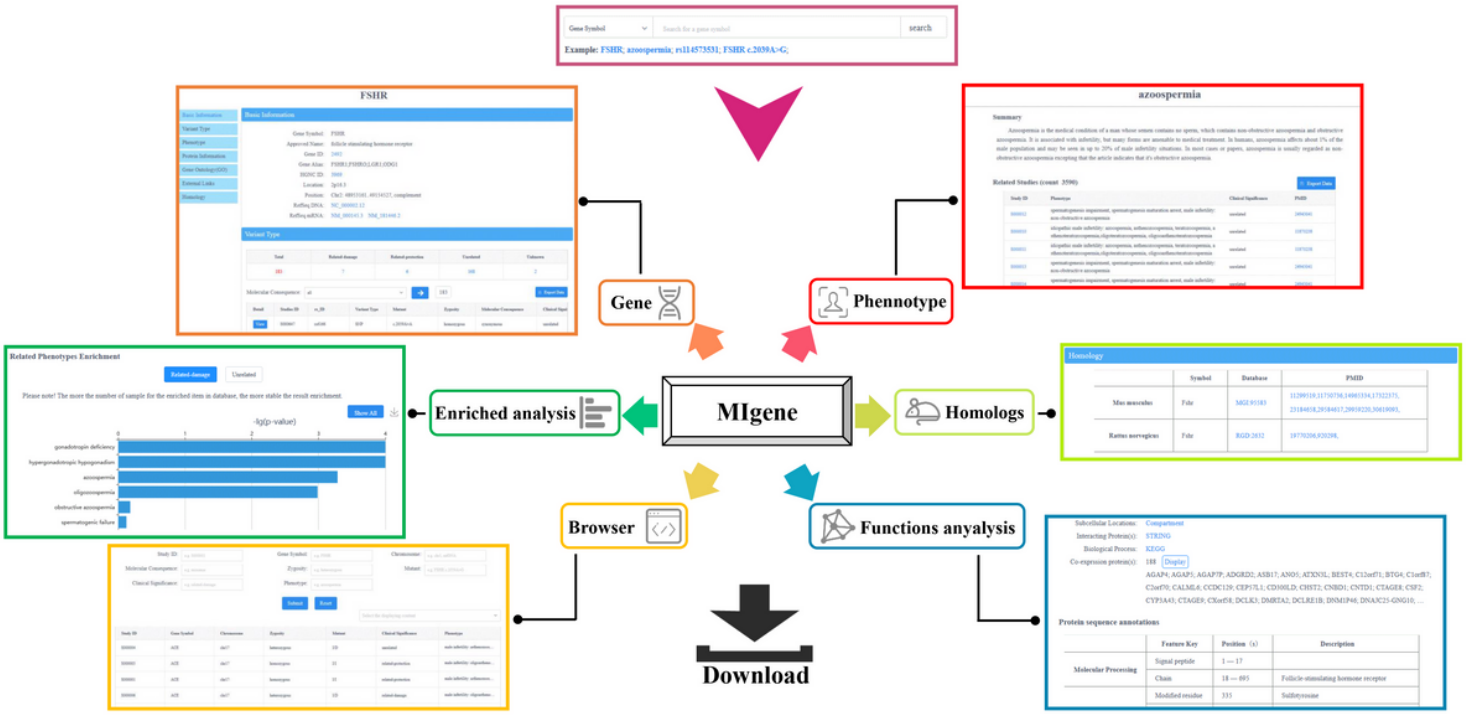

Figure 4

The functional modules and screenshot of MIgene database. The MIgene database contains 4 searching ways, 6 main modules and customized downloads.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.xls
- SupplementaryTableS2.xls
- SupplementaryTableS3.xls
- SupplementaryTableS4.xls
- SupplementaryTableS5.xls
- SupplementaryTableS6.xls
- SupplementaryTableS7.xls
- SupplementaryTableS8.xls
- SupplementaryTableS9.xls
- SupplementaryTableS10.xls
- SupplementaryTableS11.xls
- SupplementaryTableS12.xls
- SupplementaryTableS13.xls

- SupplementaryTableS14.xls
- SupplementaryTableS15.xls